WO 02/086491 A1

(54) Title: METHOD AND SYSTEM FOR CLASSIFYING CHROMATOGRAMS

(57) Abstract: A method and system for chromatogram analysis is disclosed. An aspect of the invention is a method for reducing each chromatogram to a data set that can be compared to another such data set, producing a comparison result that indicates the similarity or dissimilarity of the two chromatograms. The present invention provides a system and method that can be used to identify DNA sequence variations through chromatogram analysis. The present invention also provides a user interface to display results of chromatogram analysis, which quickly and efficiently illustrates which samples are dissimilar or similar to reference chromatograms.

METHOD AND SYSTEM FOR CLASSIFYING CHROMATOGRAMS

FIELD OF THE INVENTION

The present invention relates to the field of chromatography.

BACKGROUND OF THE INVENTION

Chromatography refers to a broad range of physical methods used to separate and analyze complex mixtures. The process of a chromatographic separation takes place within a chromatography column. A solvent, either a liquid or a gas depending on the type of chromatography process employed, moves through the column and carries the mixture to be separated. As the sample mixture flows through the column, its different components will adsorb to varying degrees. The differential rates of migration as the mixture moves over adsorptive materials provide separation for components in the mixture, since different components will elute from the chromatography column at different times. A detector measures the concentration/quantity of chemical or biological components that elute from the column.

A chromatogram is a chart that shows the detected quantity or concentration of various materials eluted from the column at different times. Different peaks on the chromatogram correspond to different components in the sample mixture. The size, shape, and/or position of peaks in the chromatogram can be used to help identify the various components in the mixture.

Chromatography may be used for many types of chemical/biological analysis and separation. For example, Denaturing High Pressure Liquid Chromatography (DHPLC) is routinely used to detect sequence variations in small sections of DNA. The technique is applied to samples in which a specific DNA fragment has been amplified by Polymerase Chain Reactions. The sample is analyzed by HPLC at a temperature at which the DNA fragment is close to denaturing, at which point the chromatographic behavior changes drastically depending on the thermal stability of each fragment. If the amplified DNA fragment exhibits a sequence variation, it will denature more readily than a non-variant fragment would, and the resulting chromatogram may be noticeably different. Samples without a sequence variation, or homozygous samples with a sequence variation will have complementary DNA strands and are referred to as "homoduplex". Samples which are heterozygous for a sequence variation will form thermally less stable "heteroduplexes", in which the DNA strands are slightly mismatched.

Fig. 1 shows a portion of an example chromatographic trace 102 resulting from DNA analysis. A typical DHPLC chromatogram 102 contains a main peak 104 corresponding to

1

leftover PCR reagents and PCR byproducts, followed by a region with one or more peaks corresponding to the DNA fragment analyzed. This region may be followed by one or more peaks resulting from the cleaning phase, in which the HPLC system gets ready for the next analysis. To detect sequence variations in the DNA mixture, analysis is performed upon the

5   region 106 of the chromatographic trace 102 that contains the peak corresponding to the variant gene sequence. Chromatographic trace 108 is a magnified view of chromatogram 102, showing the region of interest 106 having a peak 110. Normal DNA typically corresponds to a recognizable trace pattern in region of interest 106, while variant DNA would contain a trace pattern different from the "normal" trace pattern. Thus, the particular pattern or size of the

10   chromatographic trace in a specified region of interest can be used to separate variant DNA mixtures from normal DNA mixtures.

One approach to chromatogram classification, such as classification of chromatograms to identify the presence of sequence variations in DNA, is the "qualitative analysis" of chromatographic traces. Qualitative analysis generally refers to the analysis of chromatograms

15   based upon the type or shape of features in the chromatographic trace. A common approach to using qualitative analysis is to perform visual examination and comparison of traces to one or more reference traces. In the example of Fig. 1, the shape of peak 110 in chromatogram 108 could be visually compared to that same region of a reference chromatogram to determine if chromatogram 108 corresponds to variant DNA. However, qualitative analysis involving visual

20   examination is typically performed as a manual process that is often time-consuming and very subjective. Moreover, this type of approach is subject to a range of human errors.

An alternate approach is to employ "direct quantitative analysis" to classify chromatograms. The quantitative analysis approach may perform classification based upon, for example, the retention time, peak area, or number of peaks in a chromatogram trace. If an

25   algorithm is used to count peaks in the chromatographic trace, then comparison can be made between the chromatogram of the DNA mixture being analyzed and the chromatogram of normal DNA based upon the number of peaks appearing in a region of interest in the chromatographic trace. However, a significant drawback with the direct quantitative analysis approach is that DNA sequence variations can result in changes to peak shape, rather than to the

30   number of peaks, peak area, retention time, or other measures of direct quantitative analysis. Therefore, this approach may fail to adequately identify certain types of DNA sequence variations that affect the peak shape in chromatographic traces, particularly when DHPLC is used.

Thus, there is a need for an improved system and method to analyze chromatograms.

SUMMARY OF THE INVENTION

5   A method and system for chromatogram analysis is disclosed. An aspect of an embodiment of the invention provides a method for reducing each chromatogram to a data set that can be compared to another such data set, producing a comparison result that indicates the similarity or dissimilarity of the two chromatograms. In an embodiment, the present invention provides an automated system and method for identifying DNA sequence variations through

10   chromatogram analysis. An embodiment provides a method and system for automated qualitative analysis and classification of chromatograms. One embodiment of the present invention also provides a user interface to display results of chromatogram analysis, which quickly and efficiently illustrates which samples are dissimilar or similar to reference chromatograms.

15   An object of the present invention is to provide a novel system and method to effectively and efficiently analyze and compare chromatograms. Another object of the invention is to provide a method and system for comparing and identifying variant DNA. Yet another object of the invention is to provide an interface for presenting chromatogram analysis results. These and other objects, advantages, and features of the invention will be apparent to those skilled in the

20   art upon inspection of the specification, drawings, and claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings are included to provide a further understanding of the invention and, together with the Detailed Description, serve to explain the principles of the

25   invention.

FIG. 1 depicts an example chromatographic trace.

FIG. 2 shows a flowchart of a process for chromatogram analysis according to an embodiment of the invention.

FIGS. 3a, 3b, 3c, and 3d illustrate a method for baseline correction according to an

30   embodiment of the invention.

FIG. 4 shows a flowchart if a process for re-centering an analysis window according to an embodiment of the invention.

3

FIGS. 5a, 5b illustrate a method of re-centering an analysis window according to an embodiment of the invention.

FIG. 6a illustrates calculations for the area under a chromatographic trace.

FIG. 6b illustrates chromatogram reduction according to an embodiment of the invention.

FIGS. 7a and 7b show flowcharts of method for selecting a reference data set according to embodiments of the invention.

FIG. 8 illustrates an approach to chromatogram mapping according to an embodiment of the invention.

FIG. 9 depicts a user interface according to an embodiment of the invention.

FIG. 10 shows a system for chromatogram analysis according to an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to a method and system for classifying chromatographic traces. In the specification, the invention will be described with reference to specific embodiments. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention. For example, the invention is described with reference to analyzing chromatographic traces of DNA and with reference to identifying sequence variations in DNA samples. However, the disclosed principles of the invention are equally applicable to address other types of chromatograms and chromatographic analysis, e.g., for DNA genotyping. The specification and drawings are, accordingly, to be regarded in an illustrative rather than restrictive sense.

One embodiment of the present invention provides a method for reducing each chromatogram to a data set that can be compared to another such data set, producing a comparison result that indicates the similarity or dissimilarity of the two chromatograms. Fig. 2 shows a flowchart of a method for analysis of DNA chromatograms according to an embodiment of the invention. At step 202, the process receives the chromatogram data to be analyzed. In an embodiment, the chromatogram data comprises a chromatogram data file, in which the data file corresponds to digitized data for a chromatographic trace of a specific DNA sample mixture.

A region of interest in the chromatographic trace is identified for the chromatogram data (203). According to an embodiment of the invention directed to DNA analysis, the selected region of interest should exclude the main DNA peak corresponding to PCR Reagents and byproducts, while fully encapsulating the portion of the chromatographic trace that potentially

5   contains a peak corresponding to a variant gene sequence (e.g., the region of interest 106 in Fig. 1). The specific region of interest employed in the invention depends upon numerous factors, such as the type of biological/chemical material analyzed, the type of chromatography system employed, the type of analysis being performed, and the conditions under which chromatographic separation takes place. For example, a time range of 2 to 4 minutes may be

10   employed for certain types of DNA analysis using a dHPLC chromatography system. In many DNA chromatograms, the amplitude of the signal in the region of interest may be approximately two orders of magnitude lower than the amplitude of the main peak, corresponding to PCR Reagents and byproducts.

In an alternate embodiment, a specific region of interest in the chromatogram does not

15   have to be identified. Instead, in an embodiment, the data collection process is configured such that the collected or received chromatogram trace data only includes regions of interest to the analysis process. This may be accomplished, for example, by only recording chromatogram trace data during specified periods of the chromatography process when it is known that relevant data will be produced. Alternatively, the entire chromatographic trace is employed in the

20   present classification process without identifying any particular regions of interest.

Baseline correction is performed to normalize the chromatogram data for the identified region of interest (204). To further describe baseline correction, reference is made to Fig. 3a, which illustrates the region of interest 300 for an example DNA chromatogram. Region of interest 300 includes a peak 302. The preceding portion 304 of the chromatographic trace

25   extends from the main DNA peak to peak 302, and a trailing portion 306 of the chromatographic trace extends from the trailing edge of peak 302. Note that both the preceding portion 304 and the trailing portion 306 of the chromatographic trace in the region of interest 300 comprises a slope, shape, and area. Baseline correction is the process of isolating peak 302 from characteristics of the rest of the chromatographic trace, such as the slope, shape, or area

30   associated with the preceding portion 304 and trailing portion 306 in the region of interest 300.

Figs. 3b, 3c, and 3d graphically illustrate a process for baseline correction according to an embodiment of the invention. Referring to Fig. 3b, the first action of the baseline correction process is to extend the lowest point of the preceding portion 304 of the chromatographic trace

to the highest point of the trailing portion 306. This new extension 308 cuts through the lower portion of peak 302. Fig. 3c depicts the shape of the chromatographic trace if only portions 304, 306, and extension 308 are considered. To perform baseline correction, the shape shown in Fig. 3c is subtracted from the shape shown in Fig. 3a to generate the baseline-corrected shape 310 as shown in Fig. 3d. The original shape 312 of the uncorrected trace 302 is shown as a dotted line in Fig. 3d.

The process illustrated in Figs. 3b, 3c, and 3d can be implemented by starting with the first point in the data array corresponding to the chromatographic trace, and considering the straight lines joining this point to each of the subsequent points in the array. The baseline correction process uses the lowest line of the set to baseline-correct the data between the previous point and the point that produces the lowest line. These steps are repeated starting with this point, until the end of the array is reached. The baseline could be determined on a filtered array, e.g., a 3 or 5 point moving average. Alternatively, peak limits can be identified in the chromatographic trace and used to determine a baseline. In an alternate embodiment, baseline correction is not performed in the present classification process.

When chromatogram data is collected for a DNA sample, it is possible that errors may occur during the chromatography process. These errors may cause the recorded chromatographic trace for that particular sample to be "bad" or flawed data. Through experimentation, it is possible to identify a range of characteristics for a peak in the region of interest that corresponds to valid, non-erroneous data. Referring back to Fig. 2, a "bad data filter" could be employed to determine if a particular chromatographic trace corresponds to flawed data (206). Some of the characteristics that could be checked by a bad data filter include the height, shape, size, position, and/or slope of a peak in the region of interest. One or more threshold values could be established for the characteristics of the peak checked by the bad data filter. If the chromatographic trace contains a peak that exceeds a threshold characteristic, then the data is identified as "bad" data (208). Bad data can either be eliminated from the set or retained (210). If bad data is eliminated, then additional chromatogram data is received and the process described above is repeated for the additional data (216). If the bad data retained, then the identification of a particular set of data as being "bad" is maintained and used with respect to the selection of a reference data set, as described in more detail below.

Within the identified region of interest, it is possible that the peak may appear in different locations for different chromatograms. This may be caused, for example, by minor variations in the conditions under which chromatographic separation takes place. To facilitate

comparisons between different chromatograms, the positional data for peaks in different

chromatograms can be analyzed consistent with each other. One approach for consistent peak

analysis is to perform the analysis relative to a recognized reference position or offset for the

different chromatogram peaks. In the preferred approach, the classification process re-centers

5   the chromatogram data in the region of interest (212). Re-centering the chromatogram data

mathematically causes the "window" of data corresponding to the region of interest to be

centered around the peak location for each chromatogram. This re-centering inherently causes

the peaks position within the regions of interest to be consistent from one chromatogram to the

next.

10   Fig. 4 shows a flowchart of a process for re-centering chromatogram data according to an

embodiment of the invention. At 404, the average time for the region of interest is determined.

One approach for determining the average time is as follows:

$$\bar{t} = \left(\sum_{i=0}^{i=n} y_i^2 * t_i \right) / \sum_{i=0}^{i=n} y_i^2 \qquad \text{(Eq. 1)}$$

15

where y_i and t_i are the amplitude and time, respectively, for point i. The square of the amplitude

is adopted in this approach to minimize the influence of noise and baseline fluctuations. The

average time value is utilized to re-center the analysis region for the chromatography trace

(406). The analysis window for the chromatogram is shifted along the time axis such that

20   average time forms the center position of the shifted analysis window.

Figs. 5a and 5b graphically illustrate a process for re-centering according to an

embodiment of the invention. Referring to Fig. 5a, shown is an analysis window 502

surrounding a region of interest for an example chromatography trace 504. The average time

location for the amplitude values of trace 504 is noted by positional line 506 along the time axis

25   507. It can be seen that the average positional line 506 in trace 504 is not centered within

window 502. To re-center the analysis window, window 502 is shifted along the time axis 507

such that the time position of the average positional line 506 forms the center of window 502.

Fig. 5b shows the results of the re-centering process, in which analysis window 508 surrounding

the region of interest in trace 504 is centered around the average position 506, i.e., the portion of

30   analysis window 508 located to the left of the average position 106 approximately equals the

portion of window 508 located to the right of average position 506. The former location of

window 502, prior to the re-centering process, is shown in dotted lines.

Referring back to Fig. 2, the next step is to reduce the chromatographic trace to a standardized data set that can be compared numerically to other data sets (214). In an embodiment, this is performed by generating a data set for each chromatogram comprised of an array of values representative of, or derived from, the baseline-corrected and centered trace in the region of interest.

In a first approach, this data set is directly formed using the chromatogram data for a trace in the region of interest; that is, the data set comprises an array of $(time_n, amplitude_n)$ values, where $time_n$ represents the time at a particular point n along the time axis and $amplitude_n$ represents the amplitude of the trace at that $time_n$. A common set of n time points are used to determine this array of values for each chromatographic trace. Each data set can be compared or correlated to data sets for other chromatographic traces.

An alternate approach uses a "distribution" method to generate a data set for each chromatographic trace. In this approach, the integral of a chromatogram signal within the selected time region is calculated and plotted against a time axis. The integral of the chromatogram signal equates to the area formed by the chromatogram at a given point in time. To reduce the influence of noise, the integral of the square of the chromatogram signal in the region of interest can be calculated and plotted against the time axis, as follows:

$$Area(n) = \sum_{j=0}^{j=n} y_j^{\,2} \qquad\qquad \text{(Eq. 2)}$$

where y represents the chromatogram signal at a given point j, and area(n) is the area formed by the chromatographic trace from the starting point up to the jth point within the time region. The total area will be the area formed by the chromatographic trace within the entire time region.

This area determination of a trace for a region of interest is illustrated in Fig. 6a. The area(n) can be plotted against a time axis as shown in Fig. 6b. A set of time values is identified to divide the area into N equal slices. Fig. 6b illustrates a set of 10 slices for the plotted (time,area) values, based upon the percentage of the area for the trace at a given point in time. The time values can be refined by linear interpolation between the data points that bracket the target values. For instance, if the area up to point j at 0.36 minute is 8.84% of the total area and the area at point j+1 at 0.37 minute is 10.56% of the total area, the time value for the 10% mark is then 0.36 + 0.01 *(10-8.84)/(10.56-8.84) = 0.3667. In an embodiment, the 0% and 100% values, which are the start and end of the integrated array, are discarded. The values are

normalized so that the first data point is 0 and the last one is 1 using a linear transformation. The resulting data set represents the distribution of area in the selected region.

Under certain circumstances, the direct comparison approach may be more sensitive to the proper alignment of the data, and therefore small imperfections in the time alignment process

5   could have a significant impact in the correlation process. Thus, the distribution method may work better with data that have sharp features such as narrow peaks, while the direct comparison may work better with data that has broad features. One approach to reduce the sensitivity of the direct comparison method to time alignment is by shifting one data set relative to the other to vary their overlap and determine the optimal correlation value. This can be done instead of or in

10   addition to the Reference Time approach.

Other approaches may also be employed to reduce each chromatographic trace in the region of interest to a suitable data set. For example, the slope or curvature of the chromatogram signal, with or without averaging or smoothing, may be employed and combined to generate a data set representative of the chromatographic trace in a region of interest.

15   Referring back to Fig. 2, the process thereafter determines whether more chromatogram data needs to be processed (216). If so, then the previous process actions shown in Fig. 2 are repeated for each additional item of chromatogram data. If not, then the process selects a reference chromatogram from all of the previously processed chromatographic traces (218). The data set corresponding to the selected reference chromatogram provides a baseline that is

20   compared against data sets for all other traces being analyzed. For DNA analysis, the result of the comparisons provides an indication of whether sequence variations exist in DNA samples. Thus, in an embodiment, the reference chromatogram is selected to be the trace that corresponds most closely to characteristics indicative of "normal" DNA.

Fig. 7a shows a flowchart of a process for selecting a reference chromatogram according

25   to an embodiment of the invention. At 702, the ideal characteristic(s) of a normal trace pattern in the region of interest is identified. For example, for DNA analysis, the ideal characteristic for a normal trace pattern may comprise the chromatogram having the lowest variance in time. This variance could be represented as the width of the chromatographic features, and the reference chromatogram (or "homoduplex" reference) is the chromatogram having the simplest and

30   narrowest trace.

The entire set of baseline-corrected and centered chromatogram data is examined for the one trace that most closely matches the ideal characteristics (704), e.g., having the lowest variance in time. Variance may be calculated according to the following, in an embodiment:

$$\bar{t} = \sum_{i=0}^{i=n} yi^2 * t_i / \sum_{i=0}^{i=n} yi^2 \qquad \text{(Eq. 3)}$$

$$v = \sum_{i=0}^{i=n} yi^2 * (t_i - \bar{t})^2 / \sum_{i=0}^{i=n} yi^2 \qquad \text{(Eq. 4)}$$

where v represents the variance, and y_i and t_i are the amplitude and the time, respectively, for point i. Other approaches may be used in the invention to calculate variance. For example, equations 3 and 4 could be modified such that variance is computed based upon either the amplitude y_i or the square of amplitude y_i^2, for each point i.

In an embodiment, chromatogram data previously identified as being "bad data" is excluded from being eligible for selection as the reference chromatogram (706). The chromatogram having the closest match to the identified characteristics is selected as the reference chromatogram (708). Thus, for an embodiment of the invention for DNA analysis, the chromatographic trace having the lowest variance in time, which was not previously identified as being bad data, is selected as the homoduplex reference chromatogram.

Under certain circumstances, the chromatographic trace that most closely matches the "ideal" characteristics of a normal chromatogram may not actually be the most "normal" data set, e.g., because of abnormalities in the data set indicating best match which disguises other abnormal characteristics. For example, if the low-variance characteristic is the only criteria used to select the reference DNA chromatogram, then it is possible that a trace with an abnormally narrow peak is chosen as the homoduplex reference.

Fig. 7b shows a flowchart of an alternate approach to select a reference chromatogram. In this approach, the ideal characteristic(s) of the reference chromatogram are again identified (750), like the approach of Fig. 7a. The set of chromatogram data is scanned to identify chromatographic traces that correspond to the ideal characteristic(s) (752). As before, identified bad data is excluded from eligibility from being selected as the reference chromatogram (754). However, instead of selecting only a single chromatogram that corresponds to the ideal characteristics, multiple chromatograms that most closely match the ideal characteristics are

10

identified (756). For example, five chromatograms may be selected that most closely match the ideal characteristics. If low variance is the desired characteristic, then this step identifies N samples with the lowest variances in time. Each of the selected traces are compared with the other traces in the set, and the trace having the least dissimilarity with the other traces is selected

5   as the reference trace (758). This comparison can be performed, for example, by plotting the vector for each trace data set and calculating the offset difference between each trace vector; the trace with the least dissimilarity is identified based upon the smallest total distance to the other plotted traces.

With reference back to Fig. 2, the next step in the process is to compare each of the

10   chromatograms against the reference chromatogram (220). Any approach to comparing or correlating two sets of values may be employed to compare the chromatograms. In an embodiment, the comparison of data sets can be conducted by using a correlation coefficient which measures the cosine of the angle (i.e., similarity) between the two data sets, as follows:

$$\text{Cos}\,(A\,,B) = \frac{\displaystyle\sum_{i=0}^{i=n} A_i * B_i}{\sqrt{\displaystyle\sum_{i=0}^{i=n} A_i^2 * \sum_{i=0}^{i=n} B_i^2}} \qquad \text{"Similarity" Equation (Eq. 5)}$$

15   where A represents a first chromatogram data set and B represents a second chromatogram data set.

An alternate approach is to determine the sine value of the angle (i.e., the dissimilarity), as follows.

20     $$\text{Sin}\,(A\,,B) = \sqrt{1 - \text{Cos}^2(A\,,B)} \qquad \text{"Dissimilarity" Equation (Eq. 6)}$$

A third approach is to determine the distance between the vectors. After normalization, the data set of each chromatographic trace is unit length vector that terminates on the hypersphere of a unit diameter. If θ is the angle between the two vectors A and B, then the

25   Euclidian distance between the tips of these vectors can be expressed as $D = 2 * \sin\theta/2$. The distance can therefore be as a function of the cosine of the angle:

$$D(A, B) = 2 * \sqrt{\frac{1 - \text{Cos}(A, B)}{2}} \qquad \text{"Distance" Equation (Eq. 7)}$$

which can also be expressed as:

$$D(A, B) = \text{SQRT}(2(1 - \text{Cos}(A, B))) \qquad \text{(Eq. 8)}$$

The distance equation takes values of 0 for identical data sets and Sqrt(2) for orthogonal data sets. To make the calculations more user-friendly, the "2" can be removed from the equation to obtain a simplified distance value varying from 0 to 1, with otherwise equivalent properties:

$$d(A, B) = \text{SQRT}(1 - \text{Cos}(A, B)) \qquad \text{(Eq. 9)}$$

A distance-based correlation value can be defined as:

$$DBC(A, B) = 1 - d(A, B) \qquad \text{(Eq. 10)}$$

This equation has the usual properties of a correlation, in which it varies from "1" for identical sets to "0" for orthogonal sets.

Once all of the chromatogram samples have been compared against the reference chromatogram, the sample that is the most different can be selected as the "heteroduplex" reference. Alternatively, the heteroduplex reference can be manually selected.

The comparisons between each chromatogram sample and the reference chromatograms can be plotted and mapped (222 and 224). Fig.8 depicts a two-dimensional cluster map 800 in which each sample is represented by the values of its similarity (measured by DBC) with Reference 1 (Homoduplex) and Reference 2 (Heteroduplex). The example cluster map 800 in Fig. 8 illustrates a tight cluster of homoduplex samples 802, a well-separated cluster of heteroduplex samples 804, and some isolated points 806 that do not directly correspond with either the homoduplex or heteroduplex references.

The results of the mapping operation provide an immediate visual indication of samples that are likely to contain or not contain sequence variations. Automated procedures can be established to sort and identify samples that either contains or does not contain sequence

variations, based upon either the numeric results of the comparisons or based upon the results of the mapping operation, e.g., by establishing threshold distance values from the homoduplex or heteroduplex reference points. This highlights a particular advantage of the present invention, in which automated procedures can be performed to analyze and identify chromatograms according

5   to a set of defined criterion, rather than requiring a manual process of subjective visual examinations.

If different types of sequence variations are present, it is possible to identify multiple clusters on the map based on distance and standard deviation. Clusters around the homoduplex and heteroduplex references are homogeneous clusters, because samples in each cluster are

10   similar to the same reference. However, a cluster in the middle of the map may not be homogeneous. The fact that two samples have roughly the same correlation values with the two references does not indisputably establish that these samples are similar. Therefore, a cluster found in the middle of the map may need to "validated", which could be done by verifying that all the 2 by 2 comparisons within the cluster lead to a sufficient threshold of similarity. Thus, if

15   comparisons between any two samples within the cluster demonstrate a sufficient similarity, the cluster being analyzed can be considered valid. Otherwise, the samples in the cluster can be reprocessed by selecting new reference and repeating the above actions of comparing chromatograms and mapping the results.

Fig. 9 depicts a user interface 902 for analyzing chromatograms according to an

20   embodiment of the invention. User interface 902 comprises a cluster map portion 904 for displaying the results of comparing/correlating chromatographic traces, such as the cluster map shown in Fig. 8. The baseline-corrected and centered traces are displayed in a window 906. Raw chromatographic trace data (e.g., without re-centering) can be displayed in another window 908. A list of the chromatogram data files can be displayed in another window 910.

25   In an embodiment, user interface 902 is configurable to permit users to zoom in and out of each window. For example, each portion of cluster map 904 can be magnified to display individual mapped points. Features in cluster map 904 can be minimized to permit display of the entire cluster map. In an embodiment, selecting one or more points on cluster map 904 will display the corresponding chromatographic trace(s) in windows 906 and/or 908. As illustrated

30   in Fig. 9, multiple chromatographic traces may be overlaid above each other to allow visual comparison of the traces. The traces can be automatically selected and overlaid based upon identified selection criteria. For example, threshold values may be established to determine which points on the cluster map 904 correspond to a homoduplex cluster or a heteroduplex

13

cluster. The system may be configured to automatically display the traces for each cluster based upon user selection of the desired cluster. Other and additional criteria may be established for automated overlaying of one or more traces within the scope of the invention.

5   Fig. 10 depicts a system 1002 for chromatogram analysis according to an embodiment of the invention. System 1002 comprises a chromatography system 1004 that generates chromatogram data 1006, which is stored in a data storage device 1008. System 1002 further comprises a chromatogram analysis module 1010 pre-configured or configurable to perform some or all of the process actions of Fig. 2. Chromatogram analysis module 1010 includes a communication interface 1012 for sending and receiving data from data storage device 1008.

10   Chromatogram analysis module includes a baseline correction module 1014 to perform baseline correction actions, such as described with reference to Figs. 3a-3d. A window re-centering module 1016 performs actions to re-center chromatogram analysis windows, such as described with reference to Figs. 4 and 5. A chromatogram reduction module 1018 performs actions to reduce chromatograms to comparable data sets, such as described with reference to Figs. 6a and

15   6b. A chromatogram comparison module 1020 performs actions to compare chromatograms. A mapping and display module 1022 performs actions to map points on a cluster map. A bad data filter 1013 performs actions to identify potentially flawed chromatogram data. Each of these modules may access a memory device 1024 in chromatogram analysis module 1024. The chromatogram analysis module 1010 may communicate to a user station/display device 1026 to

20   display data on a user interface 1028.

        Some or all of the components in system 1002 of Fig. 10 or the process actions performed in the process of Fig. 2 may be implemented in hardware, in software, or as a combination of hardware and software. If implemented using hardware, any suitable hardware technology may be employed. For example, all or part of the chromatogram analysis module

25   1010 of Fig. 10 could be implemented using programmable logic devices such as a field programmable logic device ("FPGA"). If implemented using software, any suitable general purpose computer or dedicated programmable computing/processing device may be employed. The computer or computing device could comprise one or more processing units that perform specific operations executing one or more sequences of one or more instructions. The process

30   performed by the invention may be implemented, transmitted, or stored as any "computer-usable medium," which as used herein, refers to any medium that provides information or is usable by a computer or processing/computing device. Such a medium may take many forms, including, but not limited to, non-volatile, volatile and transmission media. Non-volatile media includes media

14

that can retain information in the absence of power. Volatile media includes media that can not retain information in the absence of power. Transmission media includes coaxial cables, copper wire and fiber optics, acoustic or light waves, and can also take the form of carrier waves; e.g., electromagnetic waves that can be modulated, as in frequency, amplitude or phase, to transmit

5   information signals.

Therefore, described is a method and system for classification of chromatograms. An embodiment of the invention can be used for automated analysis of chromatographic traces with qualitative analysis, since data relating to peak shape is reduced and compared to data for other chromatograms. Since qualitative analysis is employed, the automated procedures can be

10   employed for automated analysis of DHPLC chromatographic traces to detect sequence variations in DNA, which may cause changes in peak shapes rather than changes to direct quantitative measures. In the foregoing specification, the invention has been described with reference to specific embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and

15   scope of the invention. For example, various combinations of process actions/steps have been described. However, additional or alternate process actions or combinations of process actions may also be employed in the invention within the spirit and scope of the invention. Thus, the specification and drawings are to be regarded in an illustrative rather than restrictive sense.

WHAT IS CLAIMED IS:

1.   A method of classifying chromatograms, comprising:

receiving a first chromatogram data;

processing the first chromatogram data to generate a first data set;

5   receiving a second chromatogram data;

processing the second chromatogram data to generate a second data set; and

comparing the first data set and the second data set to perform chromatogram

classification.

10   2.   A computer usable medium having stored thereon a sequence of instructions which,

when executed by a processor, causes the processor to execute a method for classifying

chromatograms, said method comprising:

receiving a first chromatogram data;

processing the first chromatogram data to generate a first data set;

15   receiving a second chromatogram data;

processing the second chromatogram data to generate a second data set; and

comparing the first data set and the second data set to perform chromatogram

classification.

20   3.   The methods of claims 1 or 2 wherein the act of processing the first and second

chromatogram data comprises the acts of:

adjusting the first and second chromatogram data; and

reducing the first and second chromatogram data to the first and second data sets,

wherein the act of reducing is based upon consistent positioning across chromatograms.

25

4.   The method of claim 3, in which the acts of adjusting the first and second chromatogram

data comprise baseline correction.

5.   The method of claim 3, further comprising:

30   identifying a first chromatogram region of interest in the first chromatogram data;

identifying a second chromatogram region of interest in the second chromatogram data.

6.   The method of claim 3, in which the act of adjusting data in the first and second regions of interest comprises centering an analysis window around one or more trace features in a given region of interest.

5   7.   The method of claim 6, in which the act of centering comprises:
         determining an average time for the given region of interest; and
         centering the analysis window around the average time.

8.   The method of claim 3, further comprising:
10   filtering the first and second chromatogram data to identify bad data.

9.   The method of claim 8, in which the act of filtering is based upon criteria selected a group consisting of: peak height, peak area, peak shape, peak position, peak slope, peak size.

15 10.   The method of claim 3, in which the acts of reducing the first and second chromatogram data to the first and second data sets comprise determining arrays of data set values directly from the first and second chromatogram data.

11.   The method of claim 3, in which the acts of reducing the first and second chromatogram
20   data to the first and second data sets comprise:
         determining an integral of the first and second chromatogram data and plotting against a
     time axis;
         determining a set of time points; and
         forming arrays of data set values based upon the set of time points and corresponding
25   integral values for the set of time points.

12.   The method of claim 3, further comprising:
         selecting a reference chromatogram.

30 13.   The method of claim 12, in which the reference chromatogram is selected based upon first selecting a plurality of chromatograms having one or more identified characteristics that most closely match one or more reference characteristics, and identifying a single chromatogram within the plurality of chromatograms to be the reference chromatogram.

14.   The method of claim 12, in which other chromatograms are compared against the reference chromatogram.

5   15.   The method of claim 14, further comprising:
        mapping results of comparing the reference chromatogram against the other chromatograms.

16.   The method of claim 15, in which mapping is performed to a two-dimensional cluster
10   map.

17.   The method of claim 3, in which the act of comparing comprises determining a degree of similarity between the first and second data sets.

15   18.   The method of claim 3, in which the act of comparing comprises determining a degree of dissimilarity between the first and second data sets.

19.   The method of claim 3, in which the act of comparing comprises determining distance between vectors associated with the first and second data sets.
20

20.   The method of claim 3, in which the first and second chromatogram data relate to DNA analysis, wherein the reduced chromatogram data excludes a main DNA peak and fully encapsulate a possible sequence variation peak.

25   21.   The methods of claims 1 or 2 in which the act of processing the first chromatogram data comprises identifying a first qualitative characteristic for the first chromatogram data and the act of processing the second chromatogram data comprises identifying a second qualitative characteristic for the second chromatogram data, and in which the method further comprises automated comparison of the first qualitative characteristic of the first chromatogram data to the
30   second qualitative characteristic of the second chromatogram data to classify chromatograms.

22.   The method of claim 21, in which the first and second qualitative characteristics of the first and second chromatogram data comprises peak shape.

23.   The method of claim 21, further comprising:

baseline correction of the first and second chromatogram data.

24.   The methods of claims 1 or 2 in which the first and second chromatogram data comprise DHPLC chromatograms.

25.   The method of claim 24, in which qualitative analysis of peak shape is performed.

26.   The method of claim 24, in which the DHPLC chromatograms are classified based upon likelihood of SNP in DNA corresponding to the DHPLC chromatograms.

27.   The methods of claims 1 or 2 further comprising:

mapping the first and second chromatogram data.

28.   A system for classifying chromatograms, comprising:

a data storage device to store chromatogram data;

a communications interface adaptable to receive chromatogram data from the data storage device;

a data adjustment module to adjust the chromatogram data;

a reduction module to reduce the chromatogram data to a data set that can be compared against other chromatogram data sets; and

a comparison module to compare the data set against the other chromatogram data sets.

29.   The system of claim 28, further comprising a bad data filter.

30.   The system of claim 29, in which the bad data filter performs filtering based upon criteria selected from the group consisting of: peak height, peak area, peak shape, peak position, peak slope, peak size.

31.   The system of claim 28, in which the data adjustment module performs baseline correction for the chromatogram data in a region of interest.

32.   The system of claim 28, in which the data adjustment module centers the analysis window around one or more trace features in a region of interest.

33.   The system of claim 28, in which the reduction module determines an array of data set values directly from the chromatogram data.

34.   The system of claim 33, in which the array of data set values are formed by:
        selecting a set of time points in the first and second chromatogram data;
        determining amplitude values corresponding to the set of time points; and
        forming the arrays of data set values based upon the set of time points and their corresponding amplitude values.

35.   The system of claim 28, in which the reduction module determines an array of data set values based upon:
        determining an integral of the chromatogram data and plotting against a time axis;
        determining a set of time points;
        forming the arrays of data set values based upon the set of time points and corresponding integral values for the set of time points.

36.   The system of claim 28, implemented using one or more programmable logic devices.

37.   The system of claim 28, further comprising a mapping module to map results from the comparison module.

38.   The system of claim 37, further comprising a user interface to display results from the comparison module.
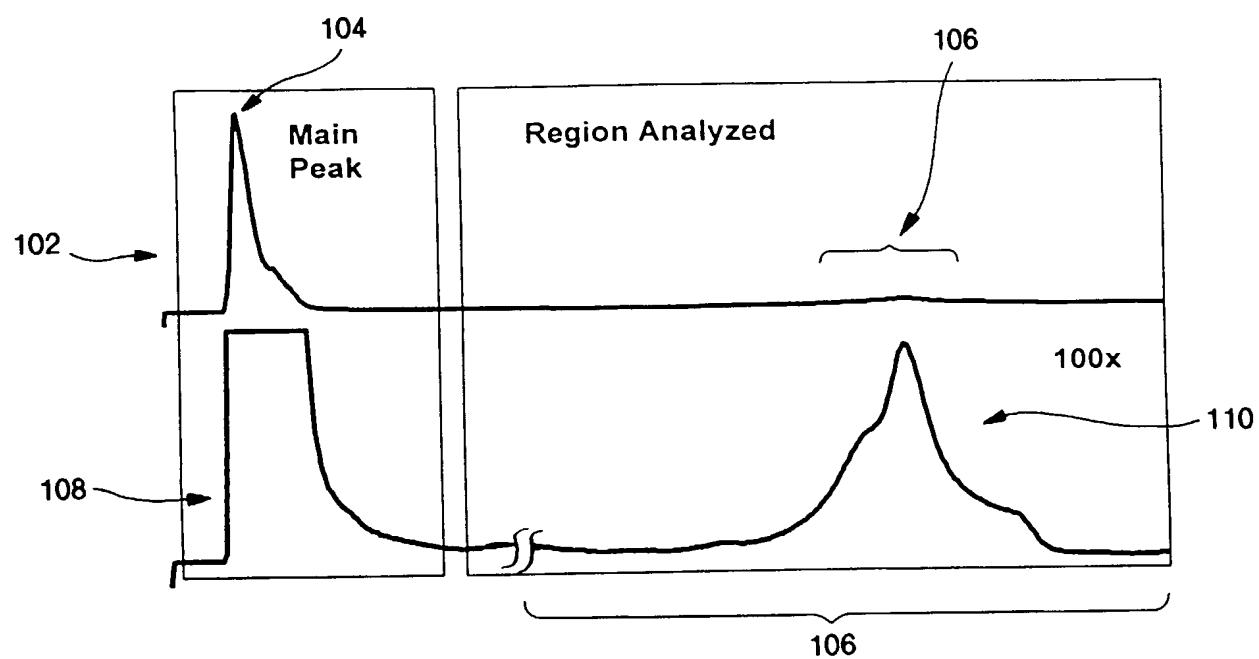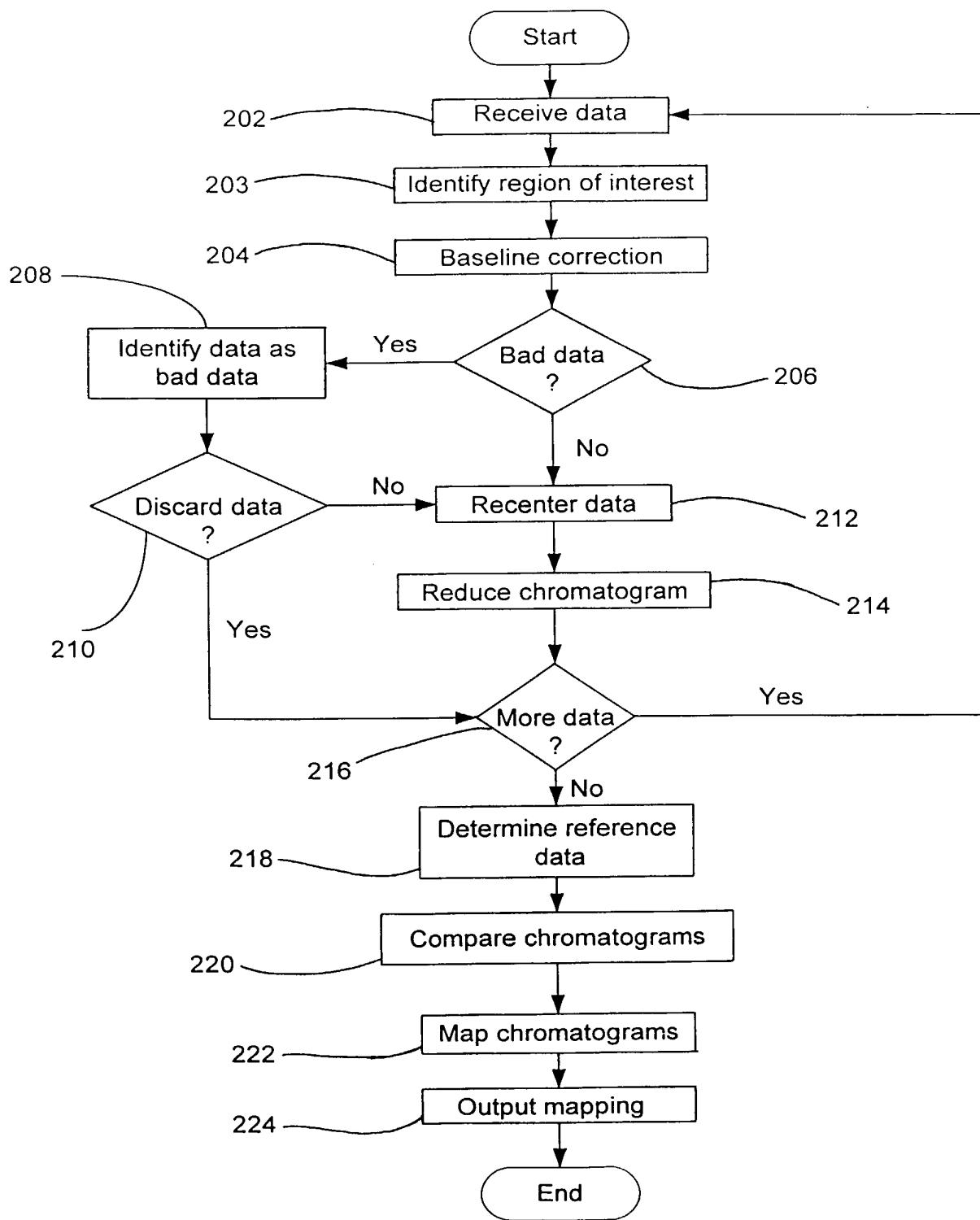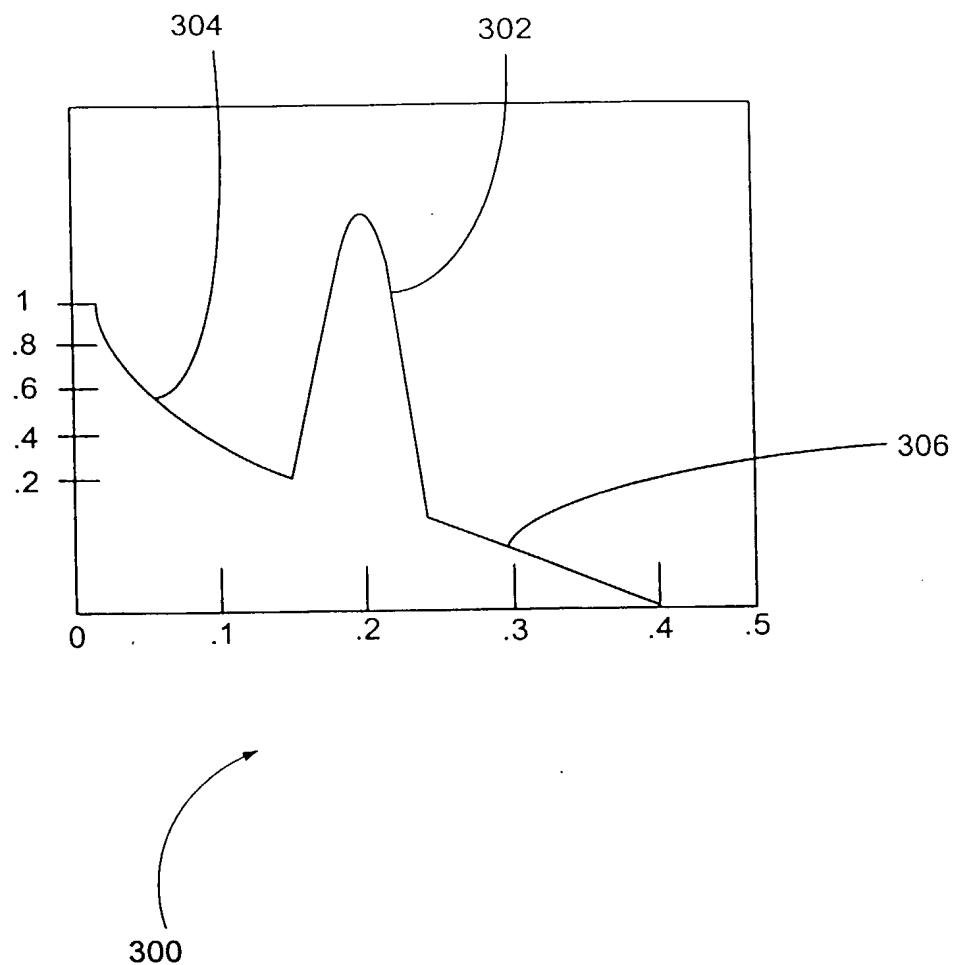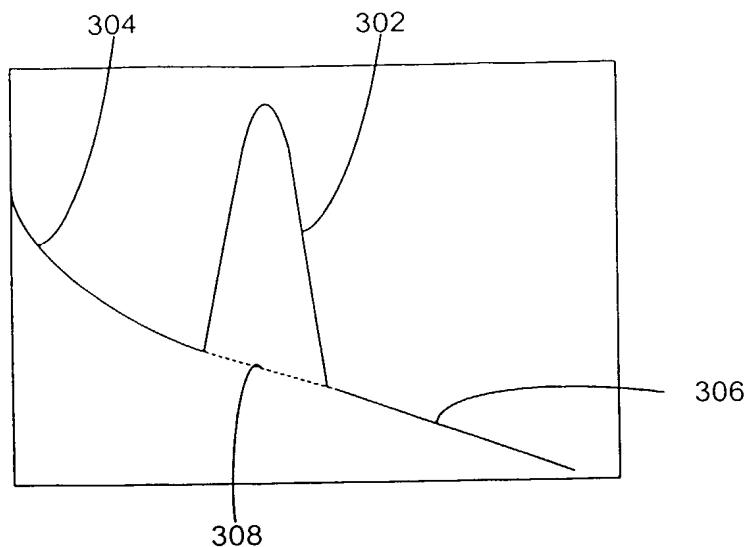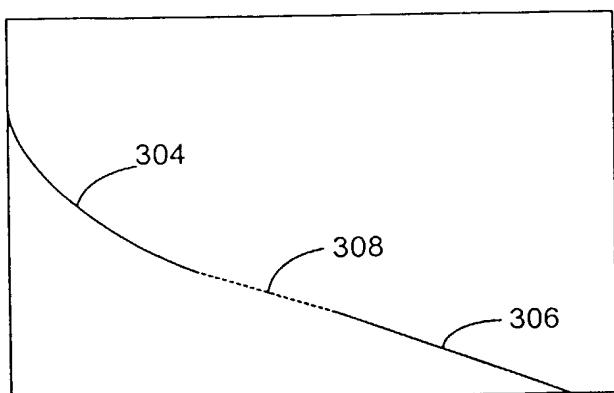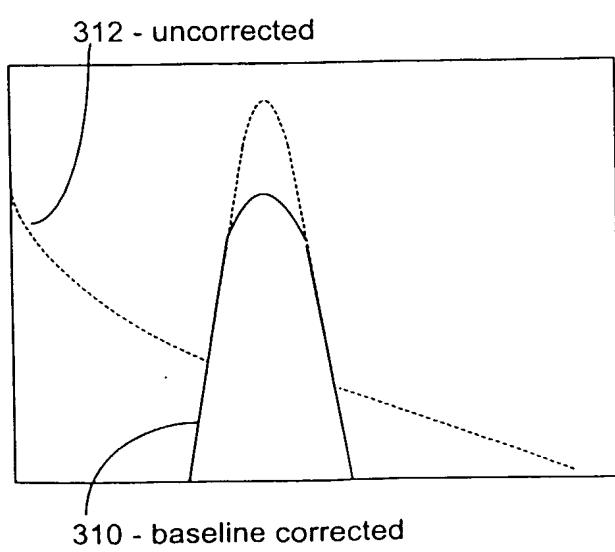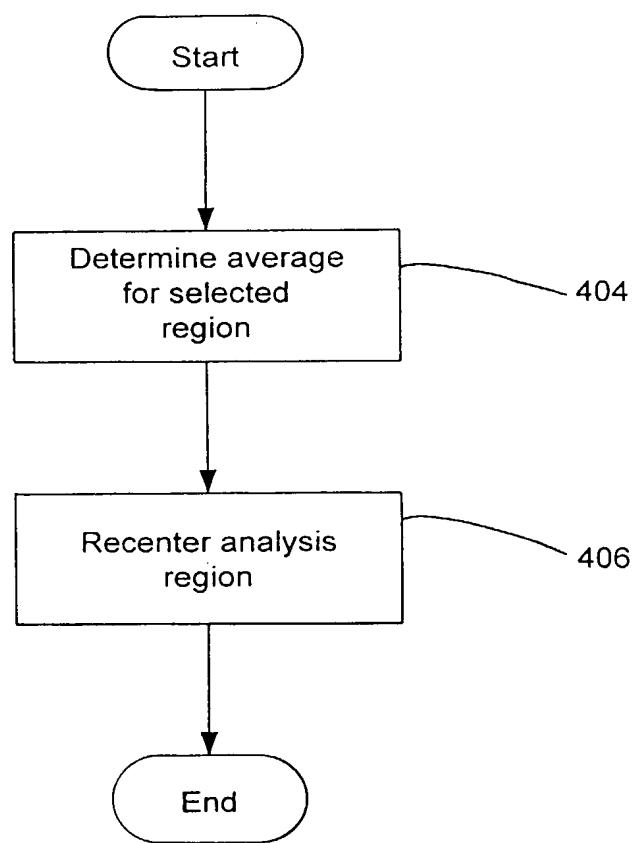
FIG. 1

2/13

FIG. 2

FIG. 3a

FIG. 3b

304   302

306

308

FIG. 3c

304

308

306

FIG. 3d

312 - uncorrected

310 - baseline corrected

FIG. 4

FIG. 5a

FIG. 5b

Quantity of materials

Chromatogram

Area(n)

the nth point

Region of interest

time

FIG. 6a

FIG. 6b

Start

Identify reference
characteristics   702

Scan data for
reference characteristics   704

Discard bad
data   706

Select data with
best match   708

End

FIG. 7a

```
        ┌─────────┐
        │  Start  │
        └────┬────┘
             │
             ▼
      ┌──────────────┐
      │  Identify reference   │────── 750
      │  characteristics      │
      └──────┬───────┘
             │
             ▼
      ┌──────────────┐
      │  Scan data for        │────── 752
      │  reference characteristics │
      └──────┬───────┘
             │
             ▼
      ┌──────────────┐
      │  Discard bad data     │────── 754
      └──────┬───────┘
             │
             ▼
      ┌──────────────┐
      │  Select set of data   │────── 756
      │  with best match      │
      └──────┬───────┘
             │
             ▼
      ┌──────────────┐
      │  Select item of data from │──── 758
      │  selected set with    │
      │  least difference     │
      └──────┬───────┘
             │
             ▼
        ┌─────────┐
        │   End   │
        └─────────┘
```

FIG. 7b

FIG. 8

FIG. 9

FIG. 10
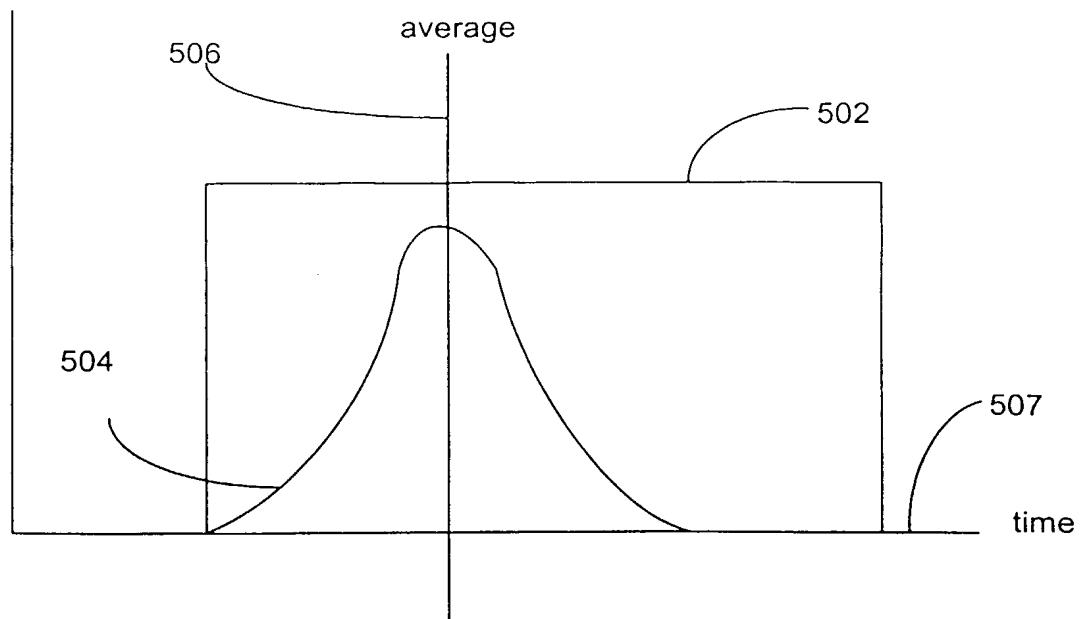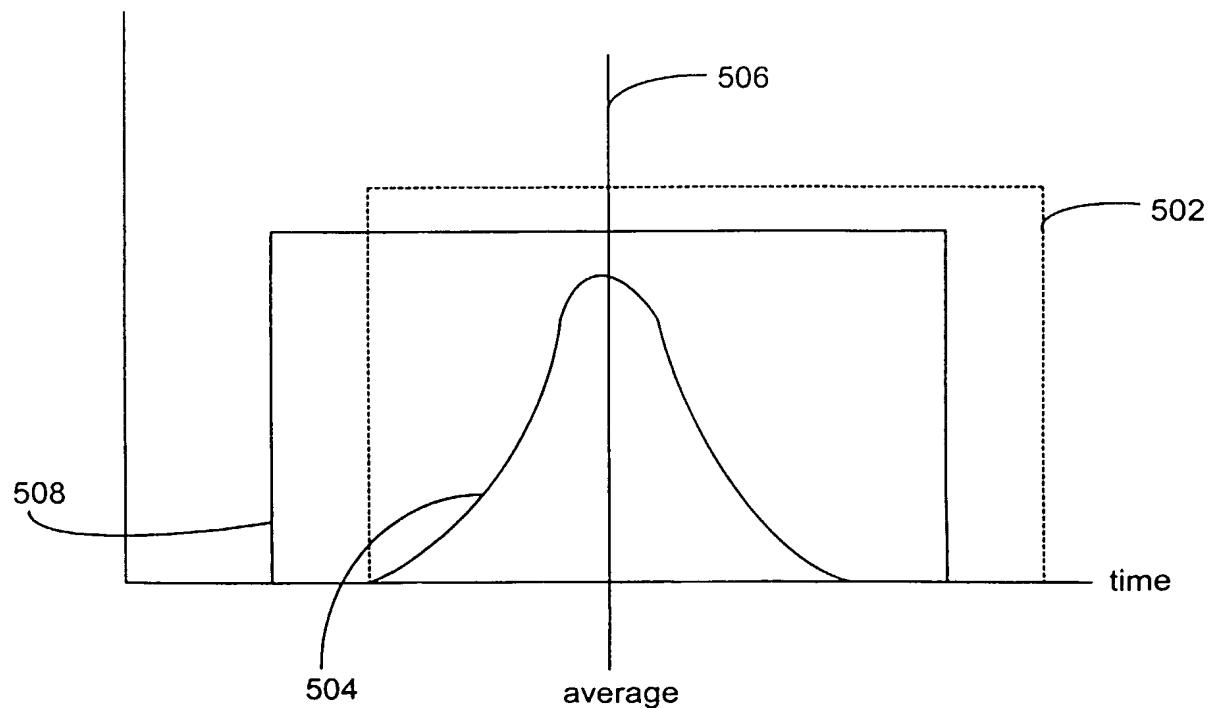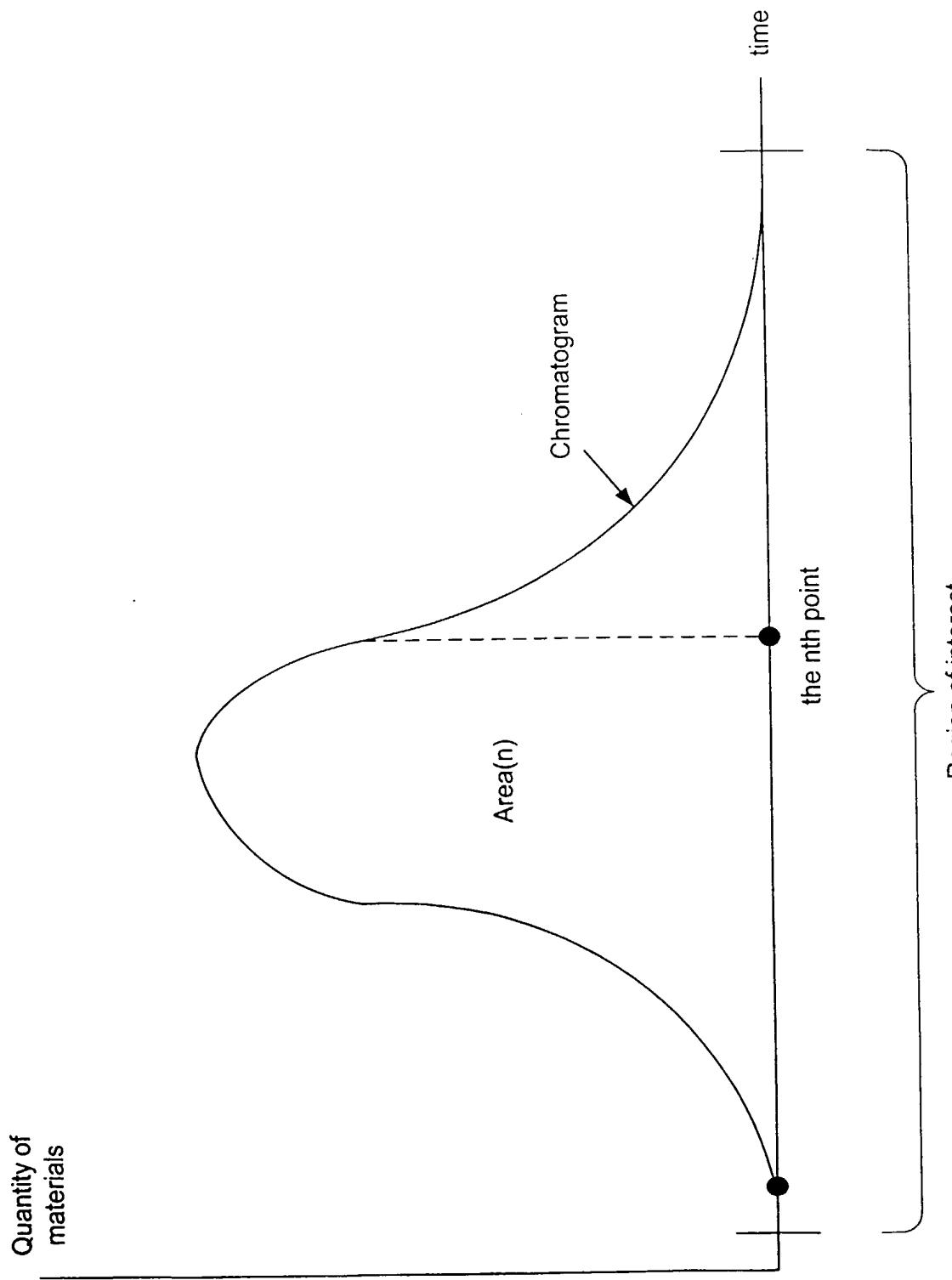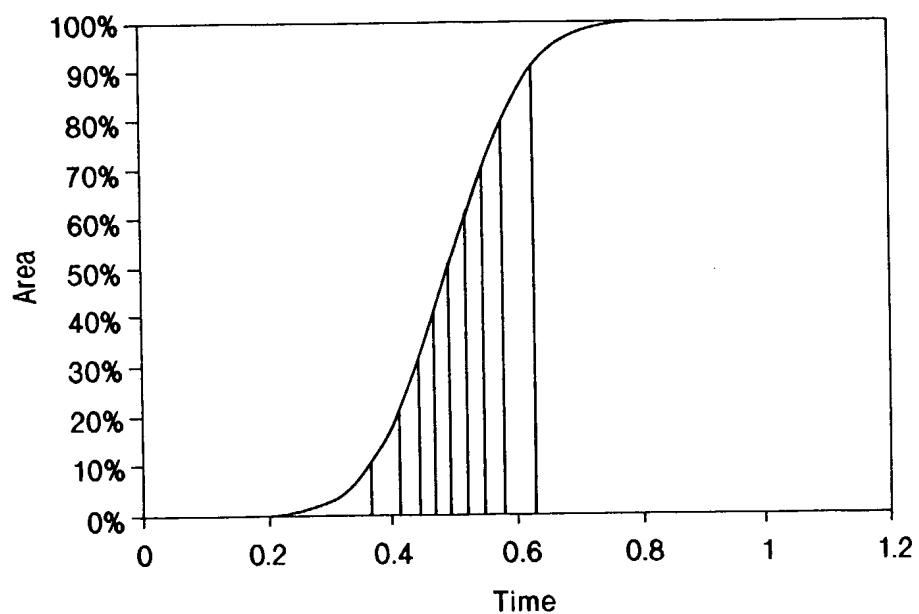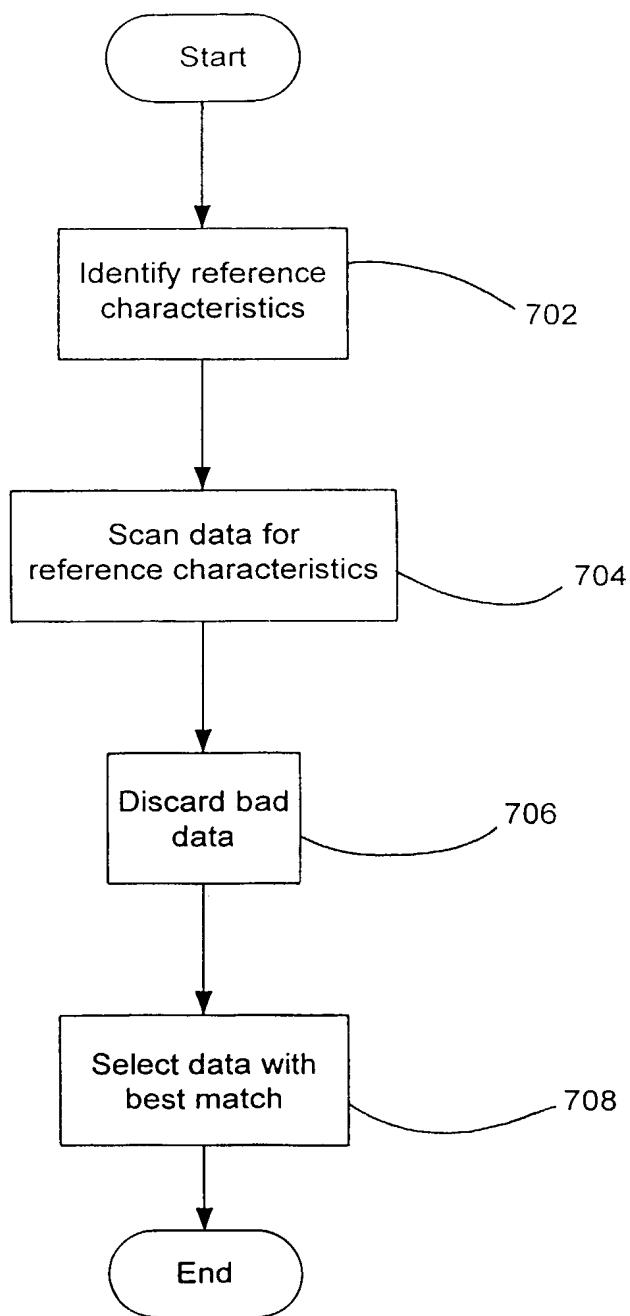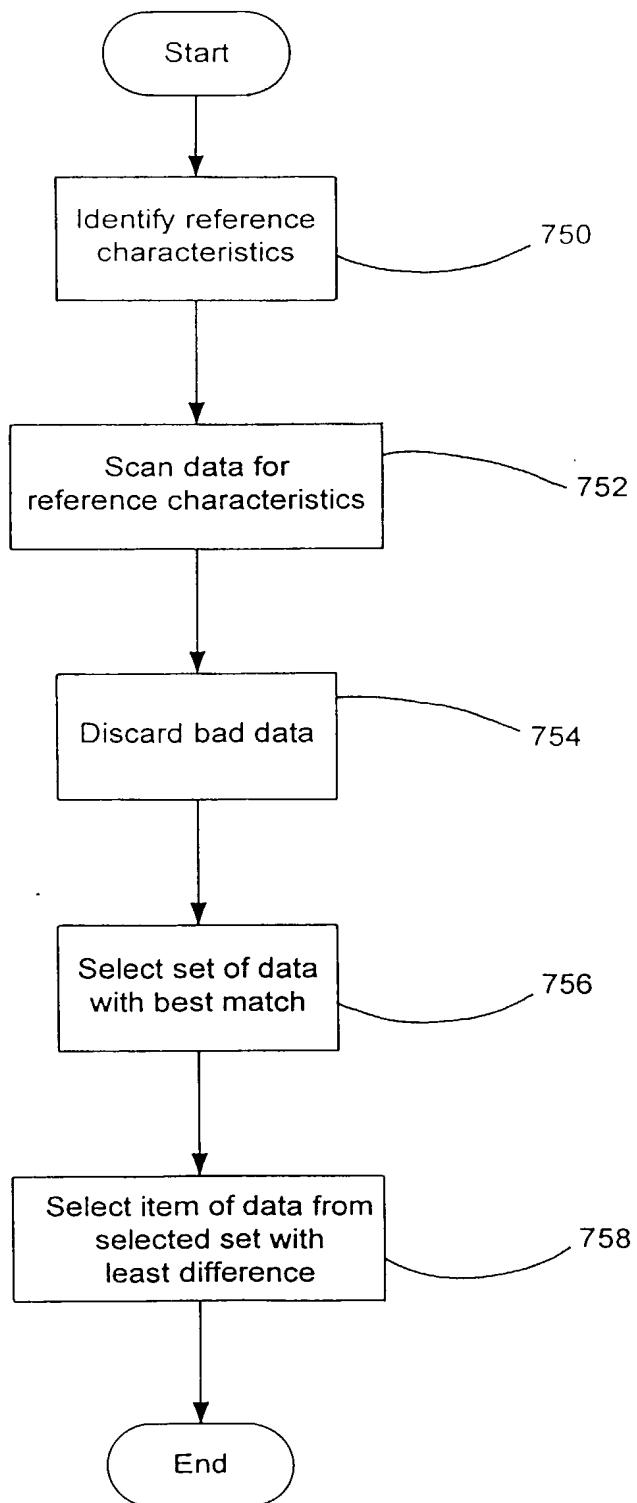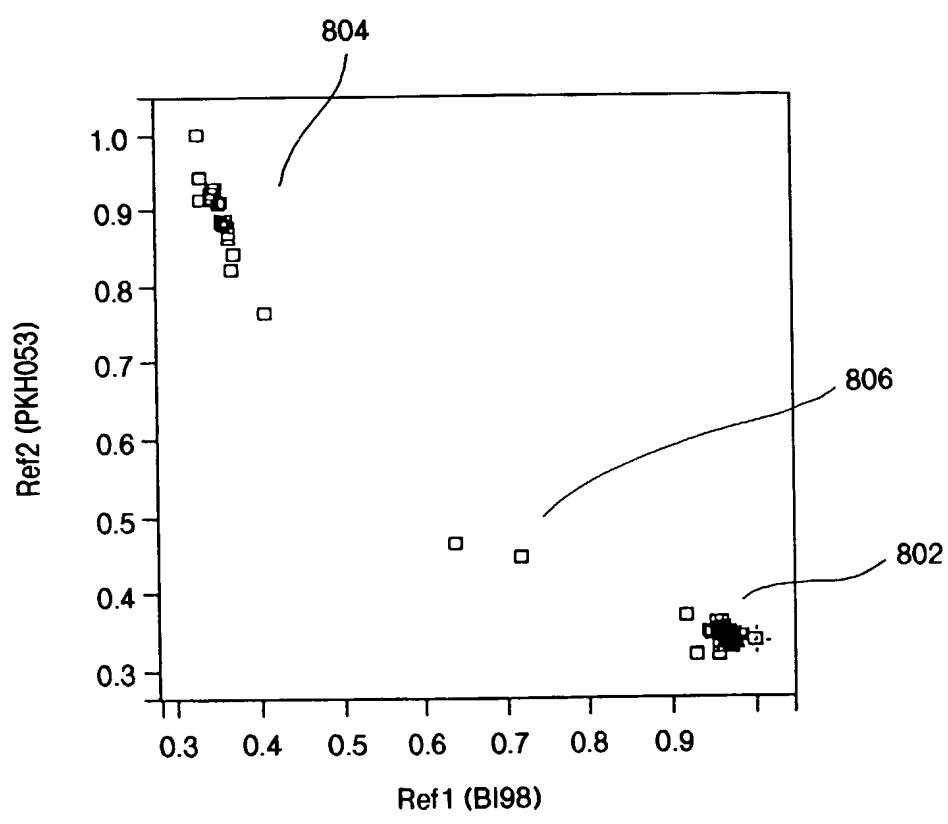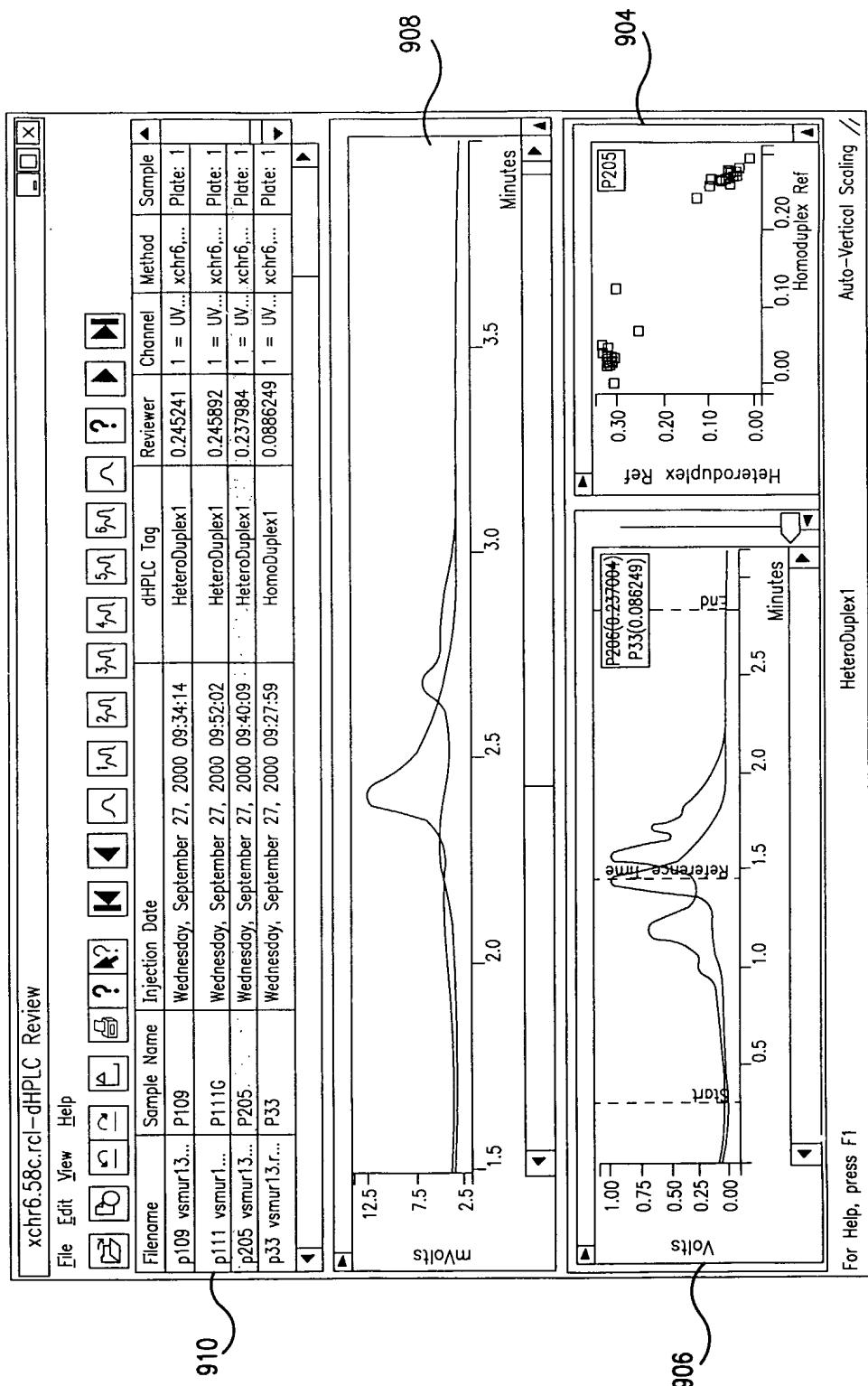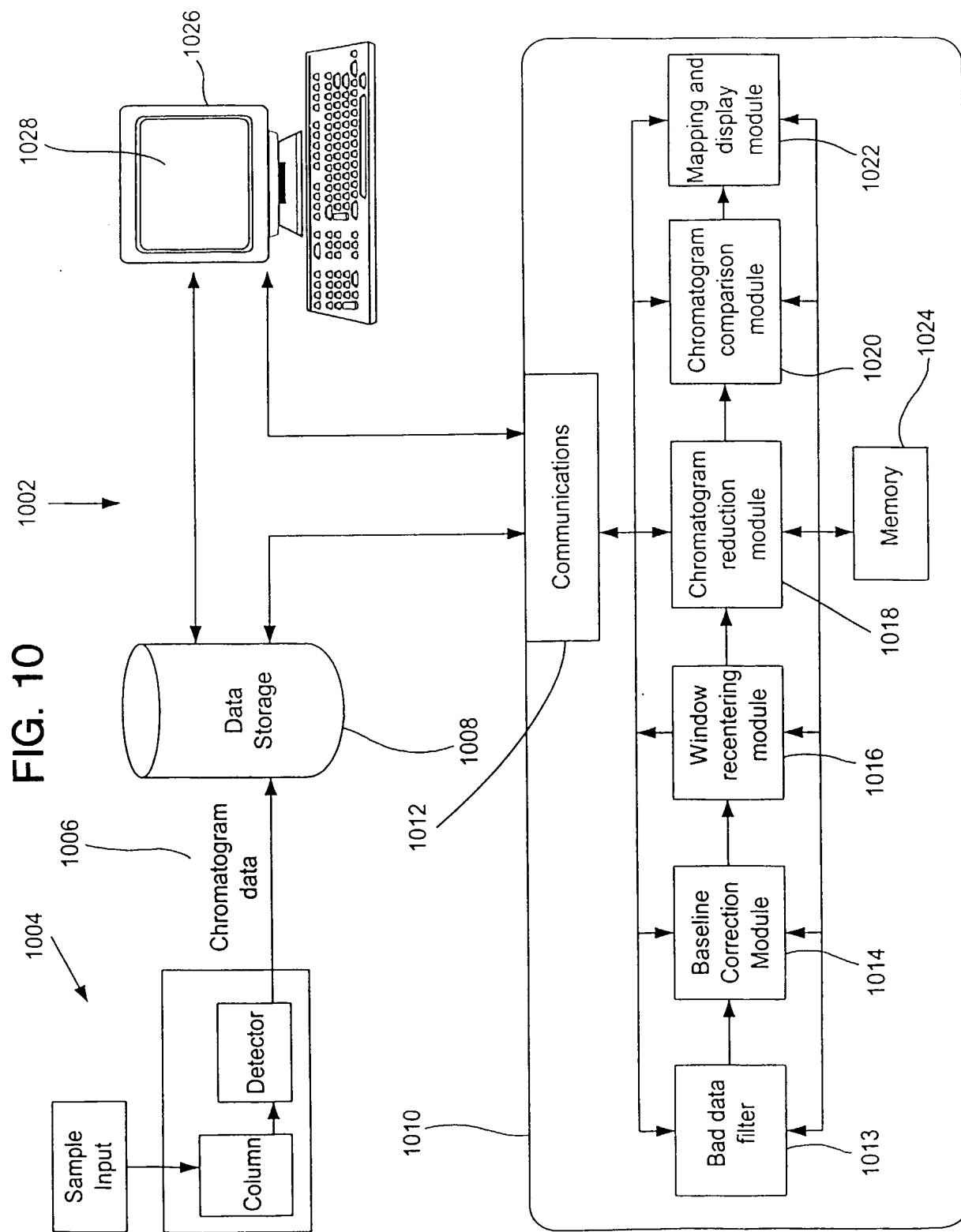
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G01N 30/86, G01N 30/88

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-INTERNAL, WPI DATA

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6195449 B1 (BOGDEN ET AL), 27 February 2001 (27.02.01)	1-38
A	US 5905192 A (WIKFORS ET AL), 18 May 1999 (18.05.99)	1-38
A	EP 0969283 A1 (HEWLETT-PACKARD COMPANY), 5 January 2000 (05.01.00)	1-38
A	EP 0296781 A2 (APPLIED BIOSYSTEMS, INC), 28 December 1988 (28.12.88)	1-38

[X] Further documents are listed in the continuation of Box C.   [X] See patent family annex.

*	Special categories of cited documents:
"A"	document defining the general state of the art which is not considered to be of particular relevance
"E"	earlier application or patent but published on or after the international filing date
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O"	document referring to an oral disclosure, use, exhibition or other means
"P"	document published prior to the international filing date but later than the priority date claimed

"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X"	document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y"	document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 July 2002	0 2. 08. 02

Name and mailing address of the International Searching Authority	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL-2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Bertil Dahl/SN Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5121443 A (TOMLINSON), 9 June 1992 (09.06.92) -- --------	1-38

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US	6195449 B1	27/02/01	NONE		
US	5905192 A	18/05/99	JP	11133015 A	21/05/99
EP	0969283 A1	05/01/00	NONE		
EP	0296781 A2	28/12/88	DE	3874235 A,T	08/10/92
			JP	1212357 A	25/08/89
			US	4837726 A	06/06/89
US	5121443 A	09/06/92	EP	0395481 A	31/10/90
			JP	2297060 A	07/12/90